Survey on Synthetic Data Generation, Evaluation Methods and GANs

Year: 2022 | Citations: 326 | Authors: Á. Figueira, Bruno Vaz

Abstract

Synthetic data consists of artificially generated data. When data are scarce, or of poor quality, synthetic data can be used, for example, to improve the performance of machine learning models. Generative adversarial networks (GANs) are a state-of-the-art deep generative models that can generate novel synthetic samples that follow the underlying data distribution of the original dataset. Reviews on synthetic data generation and on GANs have already been written. However, none in the relevant literature, to the best of our knowledge, has explicitly combined these two topics. This survey aims to fill this gap and provide useful material to new researchers in this field. That is, we aim to provide a survey that combines synthetic data generation and GANs, and that can act as a good and strong starting point for new researchers in the field, so that they have a general overview of the key contributions and useful references. We have conducted a review of the state-of-the-art by querying four major databases: Web of Sciences (WoS), Scopus, IEEE Xplore, and ACM Digital Library. This allowed us to gain insights into the most relevant authors, the most relevant scientific journals in the area, the most cited papers, the most significant research areas, the most important institutions, and the most relevant GAN architectures. GANs were thoroughly reviewed, as well as their most common training problems, their most important breakthroughs, and a focus on GAN architectures for tabular data. Further, the main algorithms for generating synthetic data, their applications and our thoughts on these methods are also expressed. Finally, we reviewed the main techniques for evaluating the quality of synthetic data (especially tabular data) and provided a schematic overview of the information presented in this paper.